·技术与应用·

DOI:10.16605/j.cnki.1007-7847.2022.10.0214

量化肿瘤浸润免疫细胞的分析方法研究进展

王建平[1,2,3], 霍子天[4], 秦 钧[2,3], 白明泽[1,2,3*]

(1. 重庆邮电大学 大数据生物智能重庆市重点实验室, 中国重庆 400065; 2. 北京蛋白质组学研究中心, 中国北京 102206; 3. 国家蛋白质科学中心(北京), 中国北京 102206; 4. 华中科技大学同济医学院附属同济医院 病理科, 中国湖北 武汉 430030)

摘 要: 免疫细胞一直被认为是维持机体平衡的重要调节因子, 对肿瘤浸润免疫细胞进行量化分析有望揭示免疫系统在肿瘤中的多方面作用。本文总结了量化肿瘤浸润免疫细胞的计算方法及相关的分析工具, 包括基于标志基因富集分析方法和反卷积分析方法开发的算法模型与工具, 列举了它们在揭示疾病发病机制以及药物治疗反应、确定肿瘤潜在生物标志物、辨别肿瘤免疫分型中的应用, 提出了它们当前存在的局限性, 并针对这些局限性问题进行了分析和展望, 以促进该领域的发展。

关键词: 肿瘤; 肿瘤浸润免疫细胞(TIC); 基因集富集分析(GSEA); 反卷积

中图分类号: Q-332, R730.3 **文献标志码**: A **文章编号**: 1007-7847(2024)02-0143-09

Advances in Analytical Methods for Quantifying Tumor-infiltrating Immune Cells

WANG Jianping[1,2,3], HUO Zitian[4], QIN Jun[2,3], BAI Mingze[1,2,3*]

(1. *Chongqing Key Laboratory of Big Data for Bio Intelligence, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*; 2. *Proteome Research Center, Beijing* 102206, *China*; 3. *National Center for Protein Sciences (Beijing), Beijing* 102206, *China*; 4. *Department of Pathology, Tongji Hospital, Tongji Medical College of Huazhong University of Science and Technology, Wuhan* 430030, *Hubei, China*)

Abstract: Immune cells have long been recognized as important regulators for the maintenance of balance in the body system. Quantitative analysis of tumor-infiltrating immune cells is expected to reveal the multi-faceted roles of the immune system in tumors. This review summarized computational methods for quantifying tumor-infiltrating immune cells from tumor tissues, including algorithm models and tools based on marker gene enrichment analysis and deconvolution method, and presented their application in revealing disease pathogenesis and drug response, potential tumor biomarkers, and cancer immunophenotyping. In addition, current limitations of these computational methods were aslo pointed out and analyzed, hoping to promote their development.

Key words: tumor; tumor-infiltrating immune cell (TIC); gene set enrichment analysis (GSEA); deconvolution

(*Life Science Research*, 2024, 28(2): 143-151)

肿瘤微环境(tumor microenvironment, TME)由肿瘤细胞、成纤维细胞、免疫细胞以及细胞外基质等组成, 其中, 肿瘤浸润免疫细胞(tumor-infiltrating immune cell, TIC)在肿瘤发生、发展和治疗中发挥着核心作用[1-2]。例如, 细胞毒性 T 淋巴细胞(又名 CD8+ T 细胞)是抗肿瘤免疫的主要效应细胞, 可以特异性地识别和杀死带有新抗原(如突变基因表达产生的肿瘤特异性抗原)的肿瘤细胞[3]。此外, 不同肿瘤类型之间的肿瘤微环境有着巨大的差异, 如 T 细胞在肝脏肿瘤和结肠肿瘤患者中的比例要显著高于其在肺部肿瘤患者中的比例[4]。因此, 有必要对肿瘤微环境进行深入探究, 其中肿瘤浸润免疫细胞的量化分析可揭示免疫细胞在肿瘤控制和治疗反应中的多方面作用。

收稿日期: 2022-10-13; **修回日期**: 2023-04-10; **网络首发日期**: 2023-06-25

作者简介: 王建平(1995—), 女, 北京人, 硕士研究生; * 通信作者: 白明泽(1982—), 男, 重庆人, 博士, 教授, 主要从事生物信息学研究, Tel: 023-62460536, E-mail: baimz@cqupt.edu.cn。

在高通量测序技术之前, 针对免疫细胞组成的研究方法主要有免疫组织化学(immunohistochemistry, IHC)、免疫荧光(immunofluorescence, IF)、流式细胞术(flow cytometry, FCM)等。这些传统的研究方法受制于其分析通量, 无法快速、大规模地量化肿瘤浸润免疫细胞的组成。得益于二代测序技术(next-generation sequencing, NGS)[5]的蓬勃发展以及检测成本的下降, 大规模描述肿瘤微环境的测序数据可以被轻松获得。在进一步的挖掘分析中, 研究人员可以基于一组免疫特异的标志基因或者免疫细胞特异表达基因矩阵计算出肿瘤浸润免疫细胞的组成。在本综述中, 我们概括了基于不同原理的量化肿瘤浸润免疫细胞的计算方法以及由此衍生的分析工具, 列举了这些算法和工具在生物学领域的应用, 总结了现有分析方法和工具的局限性, 并提出了未来在开发量化肿瘤浸润免疫细胞算法和分析工具时可能的挑战与机遇, 以期为相关领域的研究提供新的见解。

1 量化肿瘤浸润免疫细胞分析方法

1.1 基于标志基因的富集分析方法

基于标志基因的富集分析方法起源于基因集富集分析(gene set enrichment analysis, GSEA)[6], 用于评估一个预先定义的基因集是否在不同表型之间具有显著的表达差异, 其原理如图 1 所示。具体而言, 基于 GSEA 的方法首先根据表型对基因列表中的基因进行排序, 然后依据一个已知功能的基因集(如 GO 注释集、MSigDB 注释集等), 判断基因列表中的基因在已知功能基因集每条注释下的分布情况。若基因列表中的基因显著聚集在基因集的顶部或者底部, 则说明基因表达数据基于表型存在显著差异; 若基因列表中的基因随机分布在已知功能的基因集, 则认为给定的基因列表基于表型无显著差异。最后计算得出一个统计值, 也就是富集分数(enrichment score, ES), 用于表示基因的富集程度。简而言之, 给定的基因在某一个生物学过程中显著富集时则得到一个高的富集分数, 反之得到一个低的富集分数。

在量化免疫细胞浸润肿瘤细胞程度时, 基于标志基因的富集分析方法通过给定不同免疫细胞的标志基因集, 计算出免疫细胞的富集分数, 从而评估肿瘤浸润免疫细胞的程度。

基于 GSEA 原理, Barbie 等[7]开发出单样本 GSEA (single-sample GSEA, ssGSEA)算法。相较于原始的 GSEA 算法, ssGSEA 基于基因在样本中的绝对表达量对基因进行排序, 并整合基因排序的经验累积分布函数之间的差异来计算富集分数, 从而达到只有单个样本也能计算得出富集分数的目的。在此基础之上, Yoshihara 等[8]开发了 ESTIMATE (estimation of stromal and immune cells in malignant tumors using expression data)算法, 其整合了来自 11 种不同肿瘤类型的样本, 首先通过 ssGSEA 原理计算细胞基质评分和免疫评分, 从而量化浸润基质和免疫细胞的程度, 然后根据二者综合得出一个 ESTIMATE 评分, 以推断肿瘤组织中的肿瘤纯度。

此外, Becht 等[9]基于 ssGSEA 原理开发了 MCP-counter (microenvironment cell populations-counter)算法, 以对肿瘤浸润免疫细胞、成纤维细胞和上皮细胞进行量化。MCP-counter 对于每个样本都可以独立计算一个富集分数, 以评估免疫细胞及其他特定细胞类型的浸润程度, 从而使不同样本之间可以根据评分进行横向比较。xCell[10]也是基于相似的原理, 研究人员从不同的项目中提取了 489 个基因集, 整合出 64 种免疫细胞特异表达的基因集, 与 MCP-counter 相比, 其极大地丰富了算法中免疫细胞的种类。

Miao 等[11]基于 ssGSEA 算法开发了 ImmuCell-AI (immune cell abundance identifier), 这个算法可以评估 24 种免疫细胞的浸润程度, 其中包括 18 个 T 细胞亚群。基于流式细胞仪测序数据的评估结果表明, ImmuCellAI 估计 T 细胞亚群的准确度远远超过其他方法。

TIminer (tumor immunology miner)是一个综合的免疫分析算法框架[12], 包含研究人员基于 GSEA 原理开发的肿瘤免疫浸润量化分析模块, 可以评估 B 细胞、CD4+ T 细胞、CD8+ T 细胞、中性粒细胞、巨噬细胞和树突细胞这 6 种免疫细胞的浸润程度。除此以外, 该框架还集成了其他免疫分析模块, 包括人类白细胞抗原分型、新抗原预测以及免疫浸润表征。

1.2 基于部分反卷积的分析方法

反卷积模型已被广泛应用于信号和图像处理问题。在生物信息学中, 反卷积算法可以用于量化肿瘤组织样本中的免疫细胞类型和占比。基于反卷积的免疫浸润量化算法, 将肿瘤组织样本中的基因表达水平描述为存在于不同免疫细胞中的基因表达水平的线性组合[13-14]。图 2 演示了反卷积

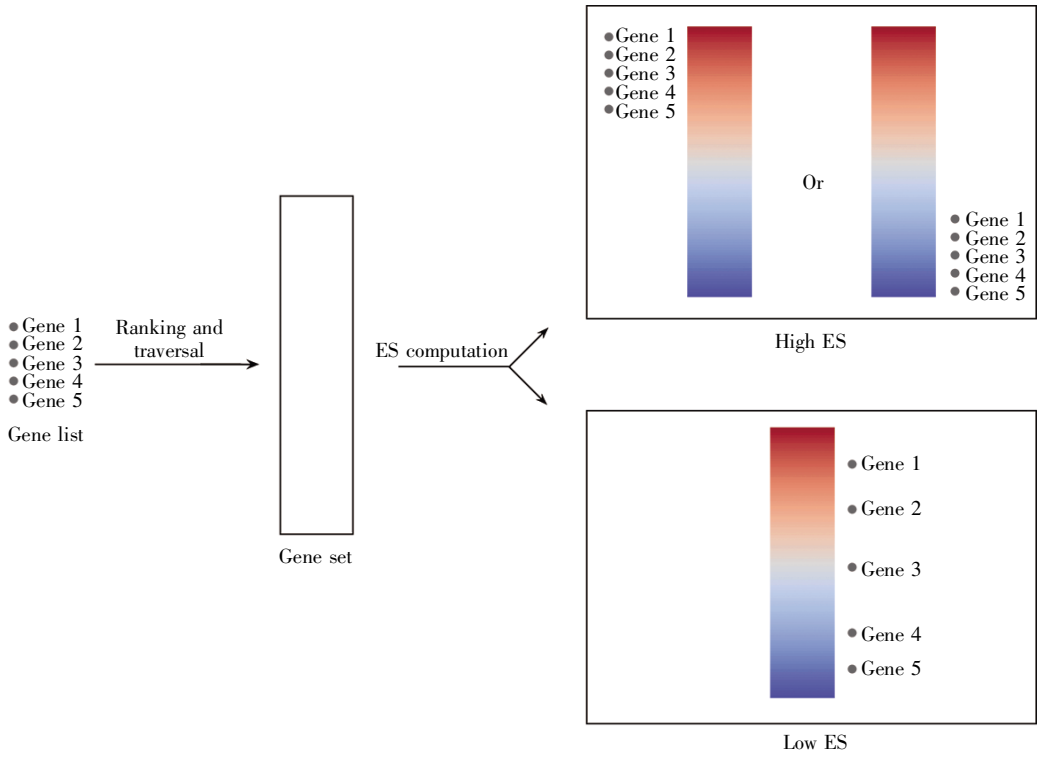


图 1　基于 GESA 的免疫浸润算法原理

给定一个基因列表并基于表型的某个指标(比如差异表达倍数)进行排序, 随后在已知不同免疫细胞类型特异表达基因的基因集中进行遍历, 根据基因在基因集中的位置分布情况计算出富集分数(ES)。若基因分布在基因集的顶端或者底部, 则代表显著上调(红色)或者显著下调(蓝色), ES 高; 若基因随机分布在基因集的各个位置, 则 ES 低。

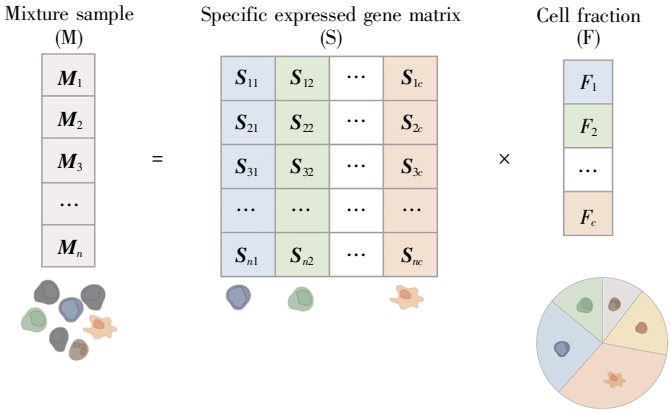
Fig.1　Immune cell infiltration algorithm based on GESA

In the GESA algorithms, the genes are ranked according to their phenotype-related measurement (such as fold-change), and then traversal is performed in immune cell specific marker gene set. The enrichment score (ES) is calculated considering the genes' position of distribution in the gene set. The ES is high when the marker genes are distributed in the top or bottom of the gene set, representing significant up-regulation (red) or down-regulation (blue), respectively, and the ES is low when genes are randomly distributed in the list.

计算免疫细胞类型和占比的过程, 首先从已知的公共数据集(如健康供体的白细胞亚群数据集)中提取不同免疫细胞类型的标志基因及其表达量, 构建出免疫细胞特征基因表达矩阵 $\boldsymbol{S}$, 然后对研究对象(即给定的肿瘤组织样本)的混合表达矩阵 $\boldsymbol{M}$ 进行反卷积运算, 求解出这些样本中不同类型的免疫细胞各自所占的比例 F。

Abbas 等[15]通过血液衍生的免疫细胞系混合物构建出免疫细胞特异性表达矩阵, 随后结合基于最小二乘回归的反卷积算法, 识别出了系统性红斑狼疮中自然杀伤(natural killer, NK)淋巴细胞和 T 辅助淋巴细胞的组成及其特异性激活模式。Qiao 等[16]提出了 PERT (perturbation model)模型, 该模型基于非负最大似然框架, 根据概率表达式进行反卷积, 并且乘以一个共享的扰动向量 $\boldsymbol{p}$, 以表征与微环境相关的细胞中基因表达的变化。与 Abbas 等[15]基于最小二乘回归的算法框架相比, 该模型对组成细胞类型的比例估计更为准确。

CIBERSORT[17] (cell-type identification by estimating relative subsets of RNA transcripts)是目前计算肿瘤浸润免疫细胞较常见的算法之一, 研究人员设计了一个由 22 种免疫细胞组成的名为“LM22”的免疫细胞特异性表达矩阵作为输入, 该算法基于线性支持向量回归进行特征选择, 自适应地选择免疫细胞特异性表达矩阵中的基因进行反卷积, 并将结果归一化为 0~1, 从而实现样本内不同细胞类型占比的量化。基于大规模、多种数据类型的验证表明, 该算法在计算具有未知含量和噪声的混合物中不同细胞类型的占比以及区分高度相关的细胞类型(如幼稚 B 细胞与记忆 B 细胞)时具有鲁棒性。此外, 该算法还支持自定义的特征表达矩阵, 研究人员可以根据感兴趣的细胞类型自定义一个特征表达矩阵来计算不同组成成分的比例。2018 年, Chakravarthy 等[18]开发了一种应



$M_j=S_{j1}F_1+S_{j2}F_2+\cdots+S_{jc}F_c$, where $j=1, 2, \cdots, n$

图 2 基于反卷积的免疫浸润算法原理[13]

对于给定肿瘤组织样本的混合表达矩阵 M, 可将其表示为由免疫细胞特征基因表达矩阵 S 和不同免疫细胞类型所占的比例 F 组成的线性方程, 即 $M=S\times F$。其中, S 为 $n\times c$ 大小的矩阵, n 为免疫细胞特征基因的数目, c 为免疫细胞类型的数目。根据公式 $M_j=S_{j1}F_1+S_{j2}F_2+\cdots+S_{jc}F_c$ 对线性方程进行求解, 得出 F。

Fig.2 Immune cell infiltration algorithm based on deconvolution[13]

For a given tumor tissue sample, the mixed expression matrix M can be expressed as a linear equation composed of the characteristic gene expression matrix S of immune cells and the proportion F of different immune cell types, that is, $M=S\times F$. In the equation, S is a $n\times c$ matrix, where n is the number of characteristic genes of immune cells, and c is the number of immune cell types. According to the formula $M_j=S_{j1}F_1+S_{j2}F_2+\cdots+S_{jc}F_c$, F can be calculated.

用于 DNA 甲基化数据的反卷积计算管道 Methyl-CIBERSORT, 其通过创建一个自定义 R 接口来生成与 CIBERSORT 共享的特征表达矩阵(其中包括 7 种不同的免疫细胞和成纤维细胞), 实现了针对 DNA 甲基化数据的免疫浸润分析。

TIMER[19] (tumor immune estimation resource)是一种多步计算方法, 其利用来自虹膜数据库的免疫特异性标志物列表和来自人类原代细胞图谱(human primary cell atlas, HPCA)的微阵列数据, 整合得到了 6 种免疫细胞的特异表达矩阵。TIMER 采用基于线性最小二乘回归方法[15]的反卷积, 并强制将所有负估计值记为零。与 CIBERSORT 不同的是, TIMER 最终的估计值没有被归一化到总和为 1, 因此, 其结果既不能直接解释为细胞组成的比例, 也不能在不同的免疫细胞类型和数据集之间进行比较。

Racle 等[20]开发了 EPIC (estimating the proportion of immune and cancer cells)模型。研究人员首先从黑色素瘤单细胞 RNA 测序(RNA sequencing, RNA-seq)数据[21]中整合出两种特征表达矩阵: 6 种血液循环免疫细胞类型, 或 5 种免疫细胞类型加上 2 种其他细胞(内皮细胞以及肿瘤相关成纤维细胞), 可以任选其一作为输入; 随后将最小二乘回归约束到反卷积算法中建立模型, 并且将每个样本中所有细胞组成的总和归一化为 0~1, 从而实现样本内不同细胞类型的比较。

为了避免混合物和特征矩阵之间不一致, Finotello 等[22]部署了一个完整的管道 quanTIseq, 实现了从读取数据、预处理到反卷积计算细胞组成占比的一站式功能集成。quanTIseq 管道基于约束最小二乘回归, 整合了来自 51 个纯化或富集免疫细胞类型的 RNA-seq 数据集构建特征表达矩阵。基于大量 RNA-seq 数据的测试表明, quanTIseq 估计出的细胞密度与 IHC 图像非常接近, 证明了该算法的可靠性。

Hunt 等[23]提出了一种反卷积方法 dtangle, 用于一系列 DNA 微阵列和批量 RNA-seq 数据的肿瘤浸润免疫细胞量化分析。研究人员使用公开的 11 个数据集来评估 dtangle, 结果表明, 与已发布的反卷积方法(CIBERSORT、EPIC 等)相比, dtangle 因对异常值和调整参数的选择具有鲁棒性, 并且计算速度更快, 具有良好的竞争力。裴晶晶等[24]开发了一个新的算法, 该算法首先计算每个基因在不同参考细胞中的前景信号; 然后, 将每个基因在不同样本中的最强信号拟合高斯分布, 并通过最大似然方法确定免疫细胞特异性表达的基因; 最后, 根据细胞特异性表达的基因, 采用逐步回归策略进行反卷积。结果表明, 该方法筛选出的特异表达矩阵与 CIBERSORT 和 dtangle 筛选出的特异表达矩阵高度重叠; 虽然该方法的计算

性能要略低于 CIBERSORT 和 dtangle, 但是其计算出的细胞组成占比最接近实验测得的真实值。

Zhang 等[25]开发了一个 ARIC 计算框架, 该框架采用了两步标志选择策略, 包括基于分量条件数的特征共线性消除和自适应孤立点标志去除。该策略可以系统地获得有效的标志物, 确保基于支持向量回归的算法能够稳健且精准地计算出 DNA 甲基化数据中的细胞比例。

MuSiC[26]是一种表征复杂组织中大量 RNA-seq 数据的细胞类型组成的方法。该算法首先计算不同细胞类型之间的相似性, 根据相似性筛选得到一组一致性基因, 通过对不同样本和不同细胞的一致性基因进行适当加权, 从而能够将不同细胞类型特异性基因的表达信息广泛应用于跨样本类型数据集。当应用于人类、小鼠、大鼠的胰岛和全肾表达数据时, MuSiC 优于现有其他方法, 尤其对于具有密切相关性的细胞类型, MuSiC 具有更好的鲁棒性。

大块组织的免疫微环境更为复杂, 其计算过程也更具挑战。2019 年, 开发 CIBERSORT 的团队对该算法进行了优化, 结合非负矩阵分解(non-negative matrix factorization, NMF)算法实现了大块组织中免疫细胞组成成分的估计, 并将新的算法命名为 CIBERSORTx[27]。

1.3　基于完全反卷积的分析方法

上一小节中描述的反卷积算法被称为部分反卷积算法, 而不是完全反卷积算法, 这是因为完全反卷积算法不仅可以计算不同细胞类型的相对占比, 而且可以提取表达谱的特征, 从而更好地解释数据。目前, 完全反卷积算法是基于 NMF 的算法原理进行开发的。NMF 是一种无监督算法, 其原理为对任意给定的非负矩阵 $\boldsymbol{V}$, 寻找一个非负矩阵 $\boldsymbol{W}$ 和一个非负矩阵 $\boldsymbol{H}$, 使得 $\boldsymbol{V}\approx\boldsymbol{W}\times\boldsymbol{H}$, 其中, $\boldsymbol{V}$ 是一个 $n\times m$ 的矩阵, 表示 R^n 空间中按列排列的 m 个数据; $\boldsymbol{W}$ 为 $n\times k$ 大小的矩阵, 为 R^n 空间中 k 个按列排列的 n 维基向量; $\boldsymbol{H}$ 为 $k\times m$ 的矩阵, 其每一列向量可视作 $\boldsymbol{V}$ 对应列向量的投影到 $\boldsymbol{W}$ 的坐标。NMF 的核心思想为矩阵分解, 常被应用于特征提取、语音识别、图像处理等多个领域[28-33]。

在生物信息学中, NMF 被广泛应用于从基因表达微阵列等高维数据中提取有意义信息。此外, 该算法还被应用于异质性组织样本中的生物标志物发现和基因表达谱的反卷积计算。Gaujoux 等[34]证明, 将来自细胞特异性标志基因的先验知识纳入基于 NMF 的方法可以显著改善完全反卷积的结果。Zhong 等[35]开发了一种 DSA (digital sorting algorithm)算法, 它首先找到一组在特定细胞类型中高表达的标志基因, 然后通过二次规划算法对样本中的细胞组分和标志基因同时进行反卷积求解。该算法的有效性在小鼠肝、脑和肺细胞的定量混合样本中得到了验证。Liebner 等[36]提出了一种 MMAD (microarray microdissection with analysis of differences)算法, 它既可以在细胞比例或特征基因表达矩阵已知时进行部分反卷积, 也可以在没有特征基因表达矩阵的情况下, 通过对数据集中的基因进行 K-Means 聚类来识别特异性表达的基因, 从而估计不同细胞类型组成成分的占比。

1.4　小结

基于标志基因富集分析方法开发的免疫浸润量化算法和分析工具大多是根据 ssGSEA 方法开发的(表 1), 其优势在于可以实现不同样本间的比较。在这些算法和分析工具中, xCell 和 ESTIMATE 被广泛应用。其中, ESTIMATE 不仅可以评估免疫细胞的浸润程度, 还可以计算基质细胞以及肿瘤纯度, 因而能够得到多维度的分析结果; xCell 可以评估 64 种不同免疫细胞的浸润程度, 相比于其他算法, 免疫细胞的种类更为丰富。基于反卷积开发出来的算法和工具是指在进行反卷积计算的时候会采用不同的算法策略。其中, CIBERSORT 是经典的反卷积算法之一, 因其采用了正则化, 所以算法具有很高的鲁棒性, 以至于后来开发的很多算法都与之进行比较。另外, 该算法的免疫细胞特征表达矩阵“LM22”被后续的很多研究用作参考矩阵[23, 37]。总的来讲, 表 1 总结的这些算法在测试数据集时普遍都具有良好的性能, 没有哪个算法具有明显的优势。

2　量化肿瘤浸润免疫细胞分析方法的应用

基于上述肿瘤浸润免疫细胞的计算方法, 不同肿瘤组织样本中免疫细胞的浸润程度可以被量化出来, 这对于揭示疾病的发病机制以及药物治疗反应具有重要意义。Zhang 等[38]通过 ssGSEA 和 CIBERSORT 算法对Ⅰ期非小细胞肺癌组织样本的免疫浸润情况进行了量化分析, 发现与无复发的原发性肿瘤相比, 在抗肿瘤免疫中发挥抑制作用的调节性 T 细胞在伴随复发的原发性肿瘤中占比最大, 其次是与受损细胞毒性和肿瘤细胞侵袭行为相关的 $CD56^{bright}$ NK 细胞。可见, 无复发的肿

表 1 常见肿瘤浸润免疫细胞分析方法

Table 1 Common algorithms for tumor-infiltrating immune cell assay

Type	Algorithm name	Method	Immune cell type	Publication year
GSEA	–	ssGSEA	–	2009[7]
	ESTIMATE	ssGSEA	–	2013[8]
	MCP-counter	ssGSEA	8	2016[9]
	xCell	ssGSEA	64	2017[10]
	ImmuCellAI	ssGSEA	24	2020[11]
	TIminer	GSEA	31 or 28	2017[12]
Partial deconvolution	–	Linear least squares regression	17	2009[15]
	PERT	Constrained least squares regression, non-negative matrix factorization	–	2012[16]
	CIBERSORT	Support vector regression	22	2015[17]
	MethylCIBERSORT	Support vector regression	6	2018[18]
	TIMER	Linear least squares regression	6	2017[19]
	EPIC	Constrained least squares regression	5	2017[20]
	quanTIseq	Constrained least squares regression	10	2019[22]
	dtangle	Linear mixed model	17	2019[23]
	–	Stepwise regression	3	2019[24]
	ARIC	Support vector regression	6 or 8	2022[25]
	MuSiC	Weighted non-negative least squares regression	5	2019[26]
	CIBERSORTx	Support vector regression, non-negative matrix factorization	22	2019[27]
Complete deconvolution	–	Non-negative matrix factorization	–	2012[34]
	DSA	Quadratic programming	6	2013[35]
	MMAD	Maximum likelihood estimation	–	2014[36]

瘤与伴随复发的肿瘤的免疫微环境具有显著差异, 因此, Ⅰ期非小细胞肺癌免疫微环境的改变可能会增加复发风险。Marczyk 等[39]对非洲裔美国人以及非洲裔以外的美国人的三阴性乳腺癌组织进行了分析, 并使用 TIDE (tumor immune dysfunction and exclusion)数据库和 CIBERSORTx 算法进行了免疫浸润分析, 发现间质瘤浸润淋巴细胞、程序性死亡受体配体 1 (programmed death-ligand 1, PD-L1)染色阳性、免疫相关通路显著富集在非洲裔美国人的肿瘤组织, 而非洲裔以外的美国人的肿瘤组织显著富集在代谢途径、TAM-M2 和免疫排除。总体而言, 非洲裔美国人的肿瘤组织存在更大程度的免疫浸润和炎症, 这可能使得其免疫系统反应更强, 对免疫检查点抑制剂和免疫治疗药物的响应率更高。Mick 等[40]对 238 名 2019 冠状病毒病(coronavirus disease 2019, COVID-19)、其他病毒性或非病毒性急性呼吸道疾病(acute respiratory illnesses, ARI)患者的上呼吸道样本进行了宏基因组测序, 并通过 CIBERSORTx 和人类肺细胞图谱数据集进行了免疫细胞组成成分的量化, 发现 COVID-19 患者的单核细胞、巨噬细胞和中性粒细胞比例显著降低, 杯状细胞、树突细胞和 B 细胞的比例显著增加, 表明 COVID-19 患者的先天免疫反应减弱与免疫微环境的细胞组成差异有关。

量化免疫细胞浸润程度还可以帮助鉴定疾病中的生物标志物。Zhong 等[41]发现, 和健康人相比, 系统性红斑狼疮患者显著高表达的基因富集在一些典型的自身免疫疾病相关通路, 例如Ⅰ型干扰素信号通路、先天免疫反应、炎症反应等。因此, 该研究进一步运用 CIBERSORT 估计出系统性红斑狼疮患者的免疫细胞占比, 然后根据占比计算不同免疫细胞之间的相关性, 发现记忆 B 细胞与激活的树突细胞、调节性 T 细胞与记忆 B 细胞、调节性 T 细胞与激活的树突细胞之间分别存在较强的正相关; 此外, CD8+ T 细胞与中性粒细胞、记忆 B 细胞和单核细胞的比例呈负相关。根据免疫细胞的分析结果, 结合基于机器学习的算法, 该研究最终筛选出更有说服力的疾病生物标志物。Wang 等[42]为了研究神经母细胞瘤发展的潜在机制, 通过 CIBERSORT 量化分析了 TCGA 数据集中神经母细胞瘤免疫细胞的比例, 随后对 22 种免疫细胞进行了 Kaplan-Meier 分析, 发现 CD8+ T 细胞比例高的患者生存率较低, 因而进一步确定了与 T 细胞增殖阳性调控相关的基因可作为预后生物标志物。

此外, 有研究人员通过 CIBERSORT 量化免疫细胞组成成分后, 再根据免疫细胞的比例进行聚类, 以此确定肾脏肿瘤患者不同的免疫亚型[43]。

Xu 等[44]运用 ESTIMATE 计算了宫颈癌患者样本的免疫评分(ImmuneScore)和基质评分(Stromal-Score), 然后通过建立蛋白质与蛋白质相互作用网络和单变量 Cox 回归分析, 获得了与预后相关的基因。

3　量化肿瘤浸润免疫细胞分析方法的挑战与机遇

值得注意的是, 以上研究大多是基于转录组数据, 当前基于蛋白质组学数据开发的量化肿瘤浸润免疫细胞的分析方法还存在缺失。一方面的原因是, 蛋白质组学研究相比转录组学和基因组学更难更复杂, 目前还正处于快速发展期。另一方面, 蛋白质组学的技术成本还不够低, 单细胞蛋白质组学的数据量相对较少。在当今生物信息学领域的“大数据”分析时代, 对于数据量小的分析, 其准确率、鲁棒性也会受到一定的限制。

当然, 也有基于蛋白质组学数据的肿瘤浸润免疫细胞研究。德国马克斯·普朗克生物化学研究所 Mann 团队[45]应用基于高分辨率质谱的蛋白质组学来表征 28 个处于稳定和活化状态的原代人类造血细胞群, 揭示了基本的细胞间通信结构和细胞类型之间以前未知的联系, 并且将蛋白质组学数据资源公开发布在 ProteomeXchange 平台。随后, Wang 等[46]结合这份免疫细胞数据, 自定义生成了一个免疫细胞的特征表达矩阵, 并通过将此矩阵应用于 CIBERSORT 算法, 对胶质母细胞瘤进行了免疫浸润分析, 探究了胶质母细胞瘤的免疫微环境以及免疫细胞与肿瘤之间的相互作用关系。由于该文献并未披露免疫细胞特征矩阵的具体计算方法, 因而无法评估该矩阵的性能以及普适性。此外, CIBERSORT 算法是基于转录组数据开发出来的, 该算法在蛋白质组学数据中的稳定性尚未可知。尽管如此, 该工作开辟了基于蛋白质组学数据的免疫浸润分析的先河, 为运用已有的蛋白质组学数据资源进行免疫浸润分析奠定了基础, 仍具有重大意义。

综上可知, 基于现有的蛋白质组学数据资源, 研究人员已然具有开发量化肿瘤浸润免疫细胞分析方法的可能性, 这将是一个很有前景的研究方向。近年来, 蛋白质组学技术正在飞速发展, 蛋白质组学数据也正在以比之前更大的规模产出, 面向蛋白质组学数据的生物信息学算法也正在相继被开发出来。肿瘤蛋白质组学推动了精准医学等领域的研究, 开发出基于蛋白质组学数据的免疫浸润量化分析方法将会使得该领域的研究更进一步, 并且为推动蛋白质组学与临床医学的结合提供重要的方法支撑。

4　总结与展望

目前, 研究人员开发出来了各种各样的量化肿瘤浸润免疫细胞的算法以及计算工具, 尤其是基于转录组数据开发出的量化分析方法数量众多, 这使得免疫浸润分析被应用于越来越多的研究中。尽管这些算法都具有良好的性能, 但肿瘤组织的复杂性和未知性, 依旧对算法的精度和鲁棒性提出了挑战。

首先, 不同组织的基因表达量可能会具有一定的差异, 不同组织中免疫细胞的肿瘤微环境也不尽相同。此前, 已有研究表明, 肿瘤相关免疫细胞的基因表达谱与来自血液的免疫细胞的基因表达谱有很大差异[47]。因此, 为了减少这种差异, 应收集不同组织的免疫细胞数据来构建标志基因集或者特征表达矩阵。

其次, 一些免疫细胞往往具有相似的表达谱或者共享部分特异性基因, 因而这些细胞往往会具有共线性问题。CIBERSORT 算法通过采取正则化方法来提取特征表达矩阵, 从而减弱多重共线性造成的影响。xCell 算法则通过对共享特异性基因的权重进行重新分配来减小多重共线性的影响。未来在开发相应的算法时也应考虑共线性问题并予以解决。

再次, 不同算法采用的免疫细胞参考数据集不同, 测试数据集也不同。通常, 不同工具之间的对比评估是在已知组成成分的 RNA 混合物或包含流式细胞术测量的细胞类型比例的实验数据集中进行的。然而, 这些实验数据集仅包含了少数免疫细胞。因此, 后续需要统一使用包含多数免疫细胞类型的参考集和测试集对算法进行评估比较, 以发现具有通用性与稳健性的算法模型。

总的来讲, 本文重点介绍了基于转录组的量化分析方法, 但是随着基因组学、蛋白质组学等数据的爆炸增长, 多组学的免疫浸润分析也亟待解决。此外, 针对蛋白质组学数据开发的肿瘤浸润免疫细胞算法还存在缺失, 因此该领域的研究存在着一定的机遇与挑战。未来, 大规模蛋白质组学必定会快速发展, 海量的数据也会爆炸式增长, 相应的基于蛋白质组学数据的量化分析算法也会

应运而生, 为肿瘤研究提供有力的支撑。

参考文献(References):

[1] FRIDMAN W H, PAGÈS F, SAUTÈS-FRIDMAN C, *et al*. The immune contexture in human tumours: impact on clinical outcome[J]. Nature Reviews Cancer, 2012, 12(4): 298-306.

[2] GAO S Y, HSU T W, LI M O. Immunity beyond cancer cells: perspective from tumor tissue[J]. Trends in Cancer, 2021, 7(11): 1010-1019.

[3] CHEN D S, MELLMAN I. Oncology meets immunology: the cancer-immunity cycle[J]. Immunity, 2013, 39(1): 1-10.

[4] LIU Y D, ZHANG Q M, XING B C, *et al*. Immune phenotypic linkage between colorectal cancer and liver metastasis[J]. Cancer Cell, 2022, 40(4): 424-437.e5.

[5] MARDIS E R. Next-generation DNA sequencing methods[J]. Annual Review of Genomics and Human Genetics, 2008, 9: 387-402.

[6] SUBRAMANIAN A, TAMAYO P, MOOTHA V K, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles[J]. Proceedings of the National Academy of Sciences of USA, 2005, 102(43): 15545-15550.

[7] BARBIE D A, TAMAYO P, BOEHM J S, *et al*. Systematic RNA interference reveals that oncogenic *KRAS*-driven cancers require TBK1[J]. Nature, 2009, 462(7269): 108-112.

[8] YOSHIHARA K, SHAHMORADGOLI M, MARTÍNEZ E, *et al*. Inferring tumour purity and stromal and immune cell admixture from expression data[J]. Nature Communications, 2013, 4: 2612.

[9] BECHT E, GIRALDO N A, LACROIX L, *et al*. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression[J]. Genome Biology, 2016, 17: 218.

[10] ARAN D, HU Z C, BUTTE A J. xCell: digitally portraying the tissue cellular heterogeneity landscape[J]. Genome Biology, 2017, 18: 220.

[11] MIAO Y R, ZHANG Q, LEI Q, *et al*. ImmuCellAI: a unique method for comprehensive T-Cell subsets abundance prediction and its application in cancer immunotherapy[J]. Advanced Science, 2020, 7(7): 1902880.

[12] TAPPEINER E, FINOTELLO F, CHAROENTONG P, *et al*. TIminer: NGS data mining pipeline for cancer immunology and immunotherapy[J]. Bioinformatics, 2017, 33(19): 3140-3141.

[13] 刘芳远, 冯学敏, 苏秀兰. 肿瘤浸润免疫细胞及其量化分析方法的研究进展[J]. 中国肿瘤生物治疗杂志(LIU Fangyuan, FENG Xuemin, SU Xiulan. Research progress of tumor infiltrating immune cells and their quantitative methods[J]. Chinese Journal of Cancer Biotherapy), 2021, 28(9): 933-941.

[14] FINOTELLO F, TRAJANOSKI Z. Quantifying tumor-infiltrating immune cells from transcriptomics data[J]. Cancer Immunology, Immunotherapy, 2018, 67(7): 1031-1040.

[15] ABBAS A R, WOLSLEGEL K, SESHASAYEE D, *et al*. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus[J]. PLoS One, 2009, 4(7): e6098.

[16] QIAO W L, QUON G, CSASZAR E, *et al*. PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions[J]. PLoS Computational Biology, 2012, 8(12): e1002838.

[17] NEWMAN A M, LIU C L, GREEN M R, *et al*. Robust enumeration of cell subsets from tissue expression profiles[J]. Nature Methods, 2015, 12(5): 453-457.

[18] CHAKRAVARTHY A, FURNESS A, JOSHI K, *et al*. Pan-cancer deconvolution of tumour composition using DNA methylation[J]. Nature Communications, 2018, 9: 3220.

[19] LI T W, FAN J Y, WANG B B, *et al*. TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells[J]. Cancer Research, 2017, 77(21): e108-e110.

[20] RACLE J, DE JONGE K, BAUMGAERTNER P, *et al*. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data[J]. eLife, 2017, 6: e26476.

[21] TIROSH I, IZAR B, PRAKADAN S M, *et al*. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq[J]. Science, 2016, 352(6282): 189-196.

[22] FINOTELLO F, MAYER C, PLATTNER C, *et al*. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data[J]. Genome Medicine, 2019, 11: 34.

[23] HUNT G J, FREYTAG S, BAHLO M, *et al*. dtangle: accurate and robust cell type deconvolution[J]. Bioinformatics, 2019, 35(12): 2093-2099.

[24] 裴晶晶, 余彩裙, 佘玉梅. 基于基因表达谱预测肿瘤浸润免疫细胞类型及比例的解卷积算法[J]. 云南民族大学学报(自然科学版) (PEI Jingjing, YU Caiqun, SHE Yumei. A deconvolution algorithm for predicting the type and proportion of tumor-infiltrating immune cells based on the gene expression profile[J]. Journal of Yunnan Minzu University (Natural Sciences Edition)), 2019, 28(4): 371-376.

[25] ZHANG W, XU H W, QIAO R, *et al*. ARIC: accurate and robust inference of cell type proportions from bulk gene expression or DNA methylation data[J]. Briefings in Bioinformatics, 2022, 23(1): bbab362.

[26] WANG X R, PARK J, SUSZTAK K, *et al*. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference[J]. Nature Communications, 2019, 10: 380.

[27] NEWMAN A M, STEEN C B, LIU C L, *et al*. Determining cell type abundance and expression from bulk tissues with digital cytometry[J]. Nature Biotechnology, 2019, 37(7): 773-782.

[28] YANG B, ZHANG X T, NIE F P, *et al*. Fast multi-view clustering via nonnegative and orthogonal factorization[J]. IEEE Transactions on Image Processing, 2021, 30: 2575-2586.

[29] XUAN J Y, LU J, ZHANG G Q, *et al*. Doubly nonparametric sparse nonnegative matrix factorization based on dependent Indian buffet processes[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(5): 1835-1849.

[30] CHE H J, WANG J. A nonnegative matrix factorization algorithm based on a discrete-time projection neural network[J]. Neural Networks, 2018, 103: 63-71.

[31] YANG M Y, WU G Y, ZHAO Q C, *et al*. Computational drug repositioning based on multi-similarities bilinear matrix factorization[J]. Briefings in Bioinformatics, 2021, 22(4): bbaa267.

[32] CHE H J, WANG J, CICHOCKI A. Bicriteria sparse nonnegative matrix factorization via two-timescale duplex neurodynamic optimization[J]. IEEE Transactions on Neural Networks and Learning Systems, 2023, 34(8): 4881-4891.

[33] RAHICHE A, CHERIET M. Blind decomposition of multispectral document images using orthogonal nonnegative matrix factorization[J]. IEEE Transactions on Image Processing, 2021, 30: 5997-6012.

[34] GAUJOUX R, SEOIGHE C. Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study[J]. Infection Genetics and Evolution, 2012, 12(5): 913-921.

[35] ZHONG Y, WAN Y W, PANG K F, *et al*. Digital sorting of complex tissues for cell type-specific gene expression profiles[J]. BMC Bioinformatics, 2013, 14: 89.
[36] LIEBNER D A, HUANG K, PARVIN J D. MMAD: microarray microdissection with analysis of differences is a computational tool for deconvoluting cell type-specific contributions from tissue samples[J]. Bioinformatics, 2014, 30(5): 682-689.
[37] AHMED U, GRAF J F, DAYTZ A, *et al*. Ultrasound neuromodulation of the spleen has time-dependent anti-inflammatory effect in a pneumonia model[J]. Frontiers in Immunology, 2022, 13: 892086.
[38] ZHANG S R, XIAO X, ZHU X L, *et al*. Dysregulated immune and metabolic microenvironment is associated with the post-operative relapse in stage Ⅰ non-small cell lung cancer[J]. Cancers, 2022, 14(13): 3061.
[39] MARCZYK M, QING T, O'MEARA T, *et al*. Tumor immune microenvironment of self-identified African American and non-African American triple negative breast cancer[J]. NPJ Breast Cancer, 2022, 8: 88.
[40] MICK E, KAMM J, PISCO A O, *et al*. Upper airway gene expression differentiates COVID-19 from other acute respiratory illnesses and reveals suppression of innate immune responses by SARS-CoV-2[J/OL]. medRxiv, 2020 [2022-06-30]. https://doi.org/10.1101/2020.05.18.20105171.
[41] ZHONG Y F, ZHANG W, HONG X P, *et al*. Screening biomarkers for systemic lupus erythematosus based on machine learning and exploring their expression correlations with the ratios of various immune cells[J]. Frontiers in Immunology, 2022, 13: 873787.
[42] WANG J P, XIAO D, WANG J X. A 16-miRNA prognostic model to predict overall survival in neuroblastoma[J]. Frontiers in Genetics, 2022, 13: 827842.
[43] QIN H S, WANG T C, ZHANG H. Identification of immune-related subtypes and characterization of tumor microenvironment infiltration in kidney renal clear cell carcinoma[J]. Frontiers in Genetics, 2022, 13: 906113.
[44] XU J J, HUANG Z, WANG Y S, *et al*. Identification of novel tumor microenvironment regulating factor that facilitates tumor immune infiltration in cervical cancer[J]. Frontiers in Oncology, 2022, 12: 846786.
[45] RIECKMANN J C, GEIGER R, HORNBURG D, *et al*. Social network architecture of human immune cells unveiled by quantitative proteomics[J]. Nature Immunology, 2017, 18(5): 583-593.
[46] WANG L B, KARPOVA A, GRITSENKO M A, *et al*. Proteogenomic and metabolomic characterization of human glioblastoma[J]. Cancer Cell, 2021, 39(4): 509-528.e20.
[47] SCHELKER M, FEAU S, DU J Y, *et al*. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data[J]. Nature Communications, 2017, 8: 2032.

(上接第 142 页)

[10] WU R L, ZHOU S, CHEN T, *et al*. Quantitative and rapid detection of C-reactive protein using quantum dot-based lateral flow test strip[J]. Analytica Chimica Acta, 2018, 1008: 1-7.
[11] LIN M S, ZHANG L W, TANG X H, *et al*. Predictive value of the HEART score combined with hypersensitive C-reactive protein for 30 d adverse cardiovascular events in patients with acute chest pain[J]. Emergency Medicine International, 2022, 2022: 3606169.
[12] PATHAK A, AGRAWAL A. Evolution of C-reactive protein[J]. Frontiers in Immunology, 2019, 10: 943.
[13] PLANK A C, MASCHKE J, ROHLEDER N, *et al*. Comparison of C-reactive protein in dried blood spots and saliva of healthy adolescents[J]. Frontiers in Immunology, 2021, 12: 795580.
[14] WANG G Y, WU C F, ZHANG Q, *et al*. C-reactive protein level may predict the risk of COVID-19 aggravation[J]. Open Forum Infectious Diseases, 2020, 7(5): ofaa153.
[15] CHEN N S, ZHOU M, DONG X, *et al*. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study[J]. The Lancet, 2020, 395(10223): 507-513.
[16] TAN C C, HUANG Y, SHI F X, *et al*. C-reactive protein correlates with computed tomographic findings and predicts severe COVID-19 early[J]. Journal of Medical Virology, 2020, 92(7): 856-862.
[17] VASHIST S K, CZILWIK G, VAN OORDT T, *et al*. One-step kinetics-based immunoassay for the highly sensitive detection of C-reactive protein in less than 30 min[J]. Analytical Biochemistry, 2014, 456: 32-37.
[18] BRAVIN C, AMENDOLA V. Wide range detection of C-reactive protein with a homogeneous immunofluorimetric assay based on cooperative fluorescence quenching assisted by gold nanoparticles[J]. Biosensors and Bioelectronics, 2020, 169: 112591.
[19] YANG X, SHU W X, WANG Y Q, *et al*. Turbidimetric inhibition immunoassay revisited to enhance its sensitivity via an optofluidic laser[J]. Biosensors and Bioelectronics, 2019, 131: 60-66.
[20] BUERKE M, SHERIFF A, GARLICHS C D. CRP-apherese bei akutem myokardinfarkt bzw. COVID-19[J]. Medizinische Klinik-Intensivmedizin und Notfallmedizin, 2022, 117(3): 191-199.
[21] VASHIST S K, VENKATESH A G, MARION SCHNEIDER E, *et al*. Bioanalytical advances in assays for C-reactive protein[J]. Biotechnology Advances, 2016, 34(3): 272-290.
[22] ZHANG Y T, GU D S. Prognostic impact of serum CRP level in head and neck squamous cell carcinoma[J]. Frontiers in Oncology, 2022, 12: 889844.
[23] BAYSAK E, GUDEN D S, ARICIOGLU F, *et al*. C-reactive protein as a potential biomarker in psychiatric practice: are we there yet?[J]. The World Journal of Biological Psychiatry, 2022, 23(4): 243-256.
[24] LIN Z, LIN Q, YU P L, *et al*. Performance evaluation of routine blood and C-reactive protein analysis using Mindray BC-7500 CRP auto hematology analyzer[J]. Annals of Translational Medicine, 2022, 10(10): 588.
[25] CIFTCI İ H, KOROGLU M, KARAKECE E. Comparison of novel and familiar commercial kits for detection of C-reactive protein levels[J]. World Journal of Microbiology and Biotechnology, 2014, 30(8): 2295-2298.
[26] DANG T, LI Z Y, ZHAO L Y, *et al*. Ultrasensitive detection of C-reactive protein by a novel nanoplasmonic immunoturbidimetry assay[J]. Biosensors, 2022, 12(11): 958.